AD_____

Award Number: MIPR 0EC5E5M0082

TITLE: The Application of Information Mining Technology to the Army Injury and Health Outcomes Database

PRINCIPAL INVESTIGATOR: Paul J. Amoroso

CONTRACTING ORGANIZATION: Army Medical Research Institute
Of Environmental Medicine
Natick, Massachusetts 01760-5007

REPORT DATE: October 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2001	Final (15 Feb 00 - 30 Sep 01)

4. TITLE AND SUBTITLE
The Application of Information Mining Technology to the Army Injury and Health Outcomes Database

5. FUNDING NUMBERS
MIPR 0EC5E5M0082

6. AUTHOR(S)
Paul J. Amoroso

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Army Medical Research Institute of Environmental Medicine
Natick, Massachusetts 01760-5007

E-Mail:

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

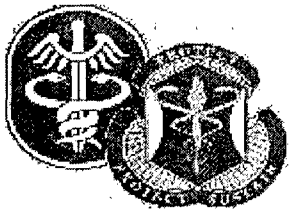
13. ABSTRACT ***(Maximum 200 Words)***

14. SUBJECT TERMS

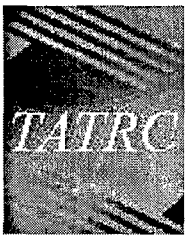
15. NUMBER OF PAGES
39

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited



DHP RFS
Final Report



The Application of Information Mining Technology to the Total Army Injury and Health Outcomes Database (TAIHOD)
Proposal Number: 1999000212

Paul J. Amoroso MD, MPH

Abstract

Problems

In the first half of the grant period we faced many challenges, principally administrative in nature, and mostly related primarily to delays in the procurement of equipment, software, and the awarding of contracts. A significant amount of training was accomplished, and while the learning curve was steep, several individuals received sufficient basic instruction to at least initiate the various processes needed to successfully accomplish the objectives of the proposal.

In addition, a special on-site block of instruction was arranged with SAS Institute. This instruction, known as a "pilot" was divided into two parts; a Warehouse Administrator (WA) pilot, and an Enterprise Miner (EM) pilot. Each pilot consisted of 5 full days of training and hands-on assistance. The WA pilot was designed to build the TAIHOD data warehouse and to provide instruction to the designated TAIHOD Warehouse Administrator (Ms. Yore). The EM pilot, was designed to walk several members of the TAIHOD staff through the process of a data mining exercise using data from the warehouse in the TAIHOD environment (Ms. Yore, Mr. Schneider, and LTC Amoroso). To get the full benefit of the mining pilot, it was necessary to select a topic suitable for demonstrating the various idiosyncrasies and capabilities of the software. We chose a practical, though arguably esoteric, exercise to examine data quality. Specifically, we attempted to identify patterns of gender misclassification across the various components of the database. This exercise had the potential utility of allowing us to find patterns of error in the database that, if corrected, could greatly improve the precision of certain analyses where there is ambiguity in the process of matching records. It also involved the use of multiple components of the database, including text (names). Several documents attached to this report provide substantial detail on this very much still-in-progress effort.

A major problem was encountered contending with the enourmously laborious task of building the SAS data warehouse. The license we obtained only allowed us to have one "administrator" for the warehouse so we focused the training on that individual (Ms. Yore). [Note: it is unlikely that even if we had a dual administrator license that we would have been able to free enough staff time to train another individual in this very complex software.]

We also discovered the accuracy of the saying that "90% of the work in any datamining effort is in data preparation step." This statement could not be more true, as a significant investment of time is required in preparing any data to be mined. This is a rather mundane task to begin with, and is therefore not met with great enthusiasm from the staff. In addition, I found that individuals with traditional training in epidemiology and biostatistics are rather resistant to the somewhat different methodology required to prepare data for mining. Most well trained epidemiologists appropriately eliminate certain variables that have no a priori likelihood of being important to an analysis. The time needed to prepare or transform certain variables for mining is further impetus to eliminate rather than incorporate the variables in a given analytic subset. Datamining requires a different mind set in that one sometimes needs to tune out preconcieved notions of what makes sense in order to give the computer the maximum chance to analyze and find associations automatically. The chances of finding something unexpected (or spurious) increases when all variables are included without any a priori assumptions to "pre-filter" the results. This of course runs very much counter to accepted scientific methods in epidemiology. Interestingly, significant resistance to anything related to "datamining" per se was also encountered among our network of PhD trained epidemiologists and biostatisticians. The very term "datamining" conjures up notions of "fishing" or "data dredging" and since these are considered bad things to traditional disciplines it was hard to solicit open minded consultation on the methodological issues encountered during the course of this very complex project. The most critical blow to this project however was certainly the loss this April of Ms. Yore. Ms. Yore had been with the TAIHOD project for almost 5 years so her relocation to another city couldn't exactly be considered premature. Nonetheless, the training and experience she gained while working on this project measurably contributed to her value in the marketplace. She departed for Atlanta in early April and with her went essentially all our expertise with warehouse administration and over half of our developed experience with Enterprise Miner. The project effectively came to a standstill weeks before her departure as we scrambled to transfer as much of her knowledge to other staff members--an endeavor in which we were only marginally successful. We arranged to keep her available as a consultant by means of a technical services contract. This at least allows us to call her with questions as they develop but this is by no means an adequte substitution for her full time effort.

Additional delays were encountered in working with our university collaborators. The first problem was the delay in getting the contract finalized. This was compounded by the fact that much of the real work on their end was to be accomplished by graduate computer science students over the summer. Once that window was missed, and the teaching commitments of the collaborators resumed, it was exceedingly difficult keeping their end of the project on track. Our inability to provide them with a sanitized dataset (i.e., sans identifiers) was also an issue since the geographic distance between our institutions is two hours by automobile. This became somewhat circular in that we needed them to write code to transform names and Social Security Numbers on the Safety data and they needed Safety data in order to test and debug their code. We ultimately had to sanitize small numbers of records by hand so they would have

something to work with but this resulted in many more iterations of the debugging process than would have been needed if we were working at the same site.

Another source of frustration, and an important piece of unfinished business in this grant, is the effective use of visualization tools. In an ideal world, one could take a dataset or series of variables and display them in various ways graphically. We discovered that Enterprise Miner's visualization toolbox is rather inadequate (a real surprise given the cost of the software). This was more than a source of disappointment however as visualization can be an especially effective tool. To overcome this shortcoming, we looked for other software packages that might do a better job in this area. Our computer science colleagues recommended DataDesk software, a relatively inexpensive package with a fairly powerful toolchest. This software is described in greater detail in the Final Result section. As a final "problem" we could point to the short-term nature of this project. It is only evident in hindsight that 12 months was an inadequate time period to accomplish the objectives of this proposal. It would be necessary to either hire an individual with considerable datamining experience at the outset, or allow greater time to develop the expertise in house. It is still our belief that once a datamining capability is fully functional, many analyses can be done much more quickly.

While having additional redundancy in who was trained to conduct the analyses required under this grant might have solved some of our problems, that option would likely be considerably more expensive. I am otherwise unsure how to retain the fully trained individuals as inevitably, once fully trained, their marketablity, in the private sector in particular can be expected to increase by 50-100%.

Deliverables

The deliverables promised for this project are described below. While the first two have not been completed, this final report will serve to fulfill the requirement for the third.

1. Report or manuscript of results from study on Risk Factors for Motor Vehicle Injuries 2. Report or manuscript of results from study on Predictors of Gulf War (Deployment) Illness. 3. A Report of the overall success of the methods developed from the study.

A methodological description of several component efforts related to development of datamining methodologies using the TAIHOD are described in Final Results (Gender missclassification.pdf and Association Rule Discovery.pdf). A chart related to Gender Misclassification is attached as (Gender Misclass Figure 1.pdf).

Reports from SAS, Inc. on the two pilot projects conducted at USARIEM are also attached (Mining Pilot summary.pdf and WA Pilot Summary.pdf) as well as a details of a contract for analyzing and improving data quality through development of automatic techniques(DATA QUALITY TAIHOD.pdf).

Note: None of these represents a finished product. These documents are included in this report because they are the best sources of information on the current status of these efforts (which are now considerably behind schedule).

The contract with UMass faculty resulted in a pledge for completion of an additional manuscript for publication in the computer science literature. A brief description of the progress of the UMass led efforts are summarized in the Final Results section. An additional "deliverable" is a license to USARIEM for UMass's Proximity software product. Paperwork for this license was received the week of 13 May 01, and will be finalized within the next several days. The attachments are as follows:

Association Rule Discovery.pdf DATA QUALITY TAIHOD.pdf Gender Misclass Figure 1.pdf Gender missclassification.pdf Mining Pilot summary.pdf UMA input.pdf

---

Expenditures

	3Q FY 00	4Q FY 00	1Q FY 01	2Q FY 01	
Element of Resource (EOR)	Apr 1 - May 31	Jun 1 - Sep 30	Oct 1 - Dec 31	Jan 1 - Mar 31	TOTALS
Travel 2100	1,965.00	4,319.00	0.00	0.00	6,284.00
Shipping 2200	0.00	0.00	0.00	0.00	0.00
Rent & Communications 2200	0.00	0.00	0.00	0.00	0.00
Contract for Services 2500	49,935.00	154,985.00	0.00	0.00	204,920.00
Supplies 2600	2,599.00	3,890.00	555.00	0.00	7,044.00
Equipment 3100	14,735.00	51,878.00	0.00	0.00	66,613.00
GRAND TOTALS	69,234.00	215,072.00	555.00	0.00	284,861.00

Financials

The final and official disbursements for this project are as follows:

EOR 21 Travel 4,534.45 25 Contract Services 123,362.48 26 Supplies/IMPAC 16,640.81 31 Equipment 64,781.98

Additionally, in an attempt to expedite the contract with the University of Massachusetts, $75,000 was "pulled back" to USAMRAA so that the contract could be negotiated from there. Therefore, the funds obligated at USARIEM represent only $210,000 of the $285,000 project award.

Management of multiple contracts was one of the major challenges to execution of this project. The one-year limitation for expenditure of P8 funds is restrictive. Furthermore, the restriction against split funding contracts from different JONOs also presents a substantial challenge.

Because it was necessary to spend this project's funds on contracts that would actually be partly funded from other sources (e.g., split funded), the EOR data provided above may not exactly reflect the actual and eventual expenditures made for the purpose of executing this study. In other words some money in support of this project was actually drawn from other fund lines and then "reimbursed" with a purchase from this project's funds to compensate.

The following items more accurately represent the breakdown of the costs of the contracts, equipment, travel, and supplies actually associated with this project:

Contract for SAS software $55,000 Support of UMass faculty $75,000 Membership in UMass CIIR $24,500 (Center for Intelligent Information Retrieval) HP Netserver computer $49,000 SAS Pilot program $25,000 SAS Enterprise Training contract $12,000 HP Color Printer $ 9,500 Travel $ 4,500 Other software $ 6,000 Contract Data Quality Improvement $ 9,500 Miscellaneous and supplies $15,000

Final Results

Multiple efforts are in mid-stream at the time of this writing. Since there are no complete products, at least with regard to the promised deliverables for this project, several attached files are included to provide background or results of several of these efforts. See attached pdf files entitled:

Association Rule Discovery.pdf DATA QUALITY TAIHOD.pdf Gender missclassification.pdf Gender Misclass Figure 1.pdf Mining Pilot summary.pdf UMA input.pdf

VISUALIZATION:

As mentioned in the Problems section, the ability to display and manipulate data graphically can be a very powerful tool for finding relationsships in data, for detecting

outliers, and for communicating results of findings to others. After acceptance of the inadequacy of the visualization tools in SAS Enterprise Miner, we explored other available tools and discovered DataDesk software.

Data Desk is a graphical data analysis package that provides fast, interactive tools for finding patterns, relationships and exceptions in data. It combines traditional statistical procedures with exploratory tools for mining and visualizing data. Data Desk is designed to work effectively with data sets all sizes - from a few hundred cases to a few million. The program optimizes computing-intensive operations resulting in speed not approached by other software running on larger, more expensive hardware. This means that Data Desk can be applied to an entire database, not just samples, and yet runs on desktop computers running Windows or Macintosh OS. Data Desk allows multiple tables (relations) to reside in the same data set. More importantly it provides seamless references between tables and for the tables themselves to change dynamically. This make it possible to start an analysis by looking at aggregated data, such as data on hospitals, then drill down to more detailed levels, examining data on physicians, and then to individual patient records. Relational database programs have these relational capabilities, but lack the statistics and graphics needed for analysis. Traditional statistics packages may provide analyses, but cannot combine information from multiple relations. Data Desk provides dynamic displays that are linked tightly together. Users can click on a point to identify it immediately and select a point or a set of points for drilldown exploration. Feedback and results are instantaneous so users are encouraged to explore the data more fully. Fast, intuitive navigational and drilldown tools are integrated intimately with sophisticated statistical modeling procedures. Databases provide fast querying tools. Statistics packages provide sophisticated statistical modeling tools. Only Data Desk provides both. Users receive the benefit of proven statistical techniques while being hidden from their complexity. Data Desk provides effective tools for developing data mining, visualization and exploration applications.

We have obtained this software and have already discovered that, as with all the other software used in this project, the learning curve is rather steep. Several plots created using survey data related to deployment of soldiers to the Gulf War have already convinced us of the utility of this software. It appears unlikely at the time of the writing of this report that we will be able to include a sample of a DataDesk plot, but a description of one application follows.

On the 1992 DMDC survey of Active Duty Servicemembers, there is a question on whether one deployed for the Gulf War. We were interested in whether or not there would be agreement between what an individual self-reported regarding their service in the Gulf War and what the DMDC deployment activation files would indicate. Errors in this widely utilized file (more than 23 published manuscripts in the peer reviewed literature have relied upon it to assess exposure) would have great bearing on the interpretation of these studies. We started by matching their responses on the question regarding service in the Gulf and the deployment activation files. We discovered discordance in as much as 15% of the cases (far greater than expected disagreement). The next question we had was what might be the common characteristics of the discordant cases? By plotting the cases against various demographic and other characteristics we quickly discovered that a high proportion of the discordant records were for people with the same arrival date in theater. If we had attempted to evaluate this data by more traditional means, we might never have been able to "see" this important attribute of these cases. Additional plots may result in even greater understanding of what is unique about these cases.

Projected Costs

Several assumptions are made in creating the estimates provided below. Based on the experiences gained in our project, it is apparent that the learning curve for use of datamining tools is quite steep. Furthermore, while there are quite a few commercial products available, they vary in terms of their quality and capabilities. All of the better packages are very costly. It is now apparent to us that having a person with substantial analytic skills working at the effort on a full time basis is essential.

Significant time must be spent in training OR, recruitment of an individual with substantial datamining experience must be done at the outset. Training capable individuals is time consuming and moderately expensive (albeit the approach we took in this project). On the other hand, recruiting a fully trained and experienced person, especially in the Northeast where individuals with SAS programming skills already command salaries in the 65-100k range, is even more expensive, but potentially more efficient.

The annual estimated/projected cost of continuing use of Datamining Methods at USARIEM is as follows:

Software costs: $ 70,000 Programmer/Analyst (contract) $135,000 Travel, training, misc. $ 10,000 Technical writing & admin support $ 27,000 Consulting (UMass faculty) $ 10,000

TOTAL $252,000

The cost of deploying these methods to other areas where large databases are in use (e.g., Army Medical Surveillance Activity) would be comparable on an annual basis. The initial year of effort would probably be equivalent to approximately 150% of the annual costs described above. Costs could be expected to go down year to year as the cost of the software comes down and/or staff can be hired as direct employees of the government.

Comments

I feel that the unrealized promises of the proposal are still achievable, but perhaps not completely so without additional resources and training time for new staff.

To engage in a project with this much potential and come up short is extremely frustrating and at times quite stressful. We have gained a great deal of capability and knowledge as a direct result of this grant, we just don't have many firm results- yet.

I think this demonstration project was a great idea, I hope that we will be afforded a way to continue to pursue the huge potential of information mining, knowledge discovery, and data quality improvement.

TATRC Scientific Review

TATRC Acquisition Review

Supporting Graphs/Charts

See Attached

The Application of Information Mining Technology to the Total Army Injury and Health Outcomes Database (TAIHOD)
Attachments to Final Report

Association Rule Discovery
Data Quality, TAIHOD
Gender Misclassification
Gender Misclassification, Figure 1
Mining Pilot Summary
UMA input
WA Pilot Summary

Association Rule Discovery (Market Basket Analysis) Project

G. Schneider (edited by LTC Amoroso)

The goal of this analysis was several fold. A DMDC survey with hundreds of responses related to demographics, military backgrounds and lifestyles, deployments, retention and career intentions, dependent and childcare issues, military compensation, benefits and programs, and family resources was used to test the utility of this function to hone in on key variables to be included in subsequent analyses. To quickly analyze all the possible associations between responses on the questionnaire, we attempted to employ the "association node" function in Enterprise Miner. While this effort is incomplete, a detailed and rather technical description of the process is detailed below.

PURPOSE

1) Implement the association rule discovery component of SAS Enterprise Miner Software to identify associations between responses from the 1992 Department of Defense Surveys of Officers and Enlisted Personnel, with an emphasis on the response that defines an individuals' deployment status for Operation Desert Shield/Desert Storm.
2) Identify problematic issues regarding the use of association rule discovery in surveillance activities and epidemiological analysis.

INTRODUCTION

This analysis was conducted by implementing the association rule discovery (market basket analysis) component of SAS Enterprise Miner software. A common use of association rule discovery is in the retail sector, which is why this methodology is often described as market basket analysis. In this common use, retailers hope to identify groups of items that customers are likely to purchase during a given trip to the store. The information obtained from these analyses can then be applied to the better development of marketing and/or advertising strategies.

Although this methodology is most commonly used in the retail sector, it has potential applications in other arenas as well. We will present the application of this tool as a means to analyze survey information, identify strengths and shortcomings of this tool for this purpose, and propose additional statistics that we feel help remedy these shortcomings.

BACKGROUND

The mathematics of association rule discovery, as it was applied in this analysis follow:

Let $J = \{R_1, R_2, \ldots\ldots, R_m\}$ denote the set of all applicable survey responses (R) included for analysis. Let D, the task relevant data, be a set of survey responses (S) where each S is a set of two or more R, such that each S is a subset of J ($S \subset J$). Each survey respondent therefore has a particular set of survey responses, S, which is a subset of all possible responses J. Furthermore, let both A and B denote subsets of J in such a manner that A denotes the left side of an association rule, and B denotes the right side of an association rule, therefore A can never equal B ($A \neq B$). Each S will contain A if and only if A is a subset of that individuals' S ($A \subset S$). Likewise, each S will contain B if and only if B is a subset of that individuals' S ($B \subset S$). An association between A and B ($A \Rightarrow B$) will therefore exist only when the following conditions are met:

The left side of the rule is included in the set of all applicable survey responses ($A \subset J$),

The right side of the rule is included in the set of all applicable survey responses ($B \subset J$),

There are individuals with both A and B ($A \cap B \neq 0$), and

A and B are not the same subset of J ($A \neq B$).

SAS Enterprise Miner outputs the following statistics for each association: support, confidence, and lift, they are defined as:

Support s, is the percentage of sets of survey responses, S, in the dataset, D, that contains both A and B, where A is associated with B ($A \Rightarrow B$).

Therefore s is a measure of prevalence:

s for $A \Rightarrow B = P(A \cap B)$, where $(A \cap B) \subset D$

Confidence c, is the percentage of sets of survey responses, S, in the dataset, D, containing A that also contains B, where A is associated with B ($A \Rightarrow B$).

Therefore c is a conditional probability:

c for $A \Rightarrow B = P(A \cap B)/P(A) = P(B \mid A)$, where $(A \cap B) \subset D$.

Lift l, is the quotient defined as the percentage of sets of survey responses, S, in the dataset, D, containing A that also contains B divided by the percentage of survey responses, S, in the dataset, D, that contains B, where A is associated with B ($A \Rightarrow B$).

Therefore l is a measure of association:

l for $A \Rightarrow B = [P(A \cap B)/P(A)]/P(B) = P(B \mid A)/P(B)$, where $(A \cap B) \subset D$.

Note that the denominator of this term P(B) is referred to as the expected confidence, which is simply the proportion of response B within the population. Therefore lift is simply the quotient of the confidence divided by the expected confidence that results in a measure of association, where:

If lift $l > 1.0$, then positive correlation,

if lift $l = 1.0$, then no correlation, and

if lift $l < 1.0$, then negative correlation.

The terms, support, confidence and lift presented above, can also be expressed via standard notation from 2x2 contingency tables. Let a, b, c, and d denote counts within the cells corresponding to: having response A and response B, having response A but not response B, not having response A but having response B, and having neither response A nor response B, respectively, as illustrated below:

Right side=B **Left side=A**	**B+**	**B-**
A+	**a** = A and B	**b** = A, not B
A-	**c** = not A, B	**d** = not A, not B

Via this notation we can demonstrate that:

$$P(A \cap B) = a/(a+b+c+d)$$

Therefore, support *s* is analogous to the epidemiological term prevalence, where prevalence is the proportion of the total population with both response A and response B,

$$P(B \mid A) = a/(a+b)$$

Therefore, the conditional probability termed confidence *c* is analogous to the epidemiological term cumulative incidence, where cumulative incidence is the proportion with response B among those with response A. Lastly, by dividing this term by the probability of having response B in the entire population,

$$P(B \mid A) / P(B) = [a/(a+b)] / [(a+c)/(a+b+c+d)]$$

The measure of lift is obtained.

Notice that these statistics are focused on the upper row of the 2x2 table. Thus, the prevalence of people with both response A and response B, and the cumulative incidence

of response B among those with response A are presented as via the terms support and confidence, respectively. We felt it necessary to also calculate the cumulative incidence without response A, but with response B.

$$P(B \mid A_0) = c/(c+d)$$

This enabled the calculation of a cumulative incident ratio, a measure of relative risk (RR), for each association.

$$RR = P(B \mid A) / P(B \mid A_0) = [a/(a+b)] / [c/(c+d)]$$

where:

If RR > 1.0, then response B was more prevalent among those with response A than those without response A,

if RR = 1.0, then response B was equally prevalent among those with and without response A, and

if RR < 1.0, then response B was less prevalent among those with response A than those without response A.

METHODS

We implemented the association rule discovery component of SAS Enterprise Miner Software to identify all two-way, three-way and four-way associations, with a confidence of greater than 10%, between 569 possible responses that originated from 52 of 138 questions from the 1992 Department of Defense Survey of Officers and the same 52 of

139 questions from the 1992 Department of Defense Survey of Enlisted Personnel. A total of 13,734 surveys were included in this analysis. Particular emphasis was placed on the response that defined an individuals' self-reported deployment status for Operation Desert Shield/Desert Storm.

For each association, we calculated the RR, and corresponding 99.0% confidence intervals, of responding to the right hand side response (denoted as B above), among those with and without the left hand side response (denoted as A above). Realizing that there would likely be an enormous quantity of associations, those that were not significant at the p=0.01 level were discarded. Additionally, the results were grouped by the number of responses in the association (2-way, 3-way, or 4-way) and then sorted by descending RR values. Samples of associations with the largest values of RR are presented for each n-way association. This process was also conducted for associations in which the left hand response denoted that an individual was deployed for Operation Desert Shield/Desert Storm.

RESULTS

Nearly 92 million associations were identified, 100,731 2-way, 5,335,949 3-way and 86,486,408 4-way associations.

The association results are not yet available.

Note: An emphasis should be placed on 2-way associations with "deployed" in the left hand column

DISCUSSION

Association rule discovery is a powerful tool. Although not provided in its output, enough information is provided that the ratio of the prevalence of the right side response in an association, between those with and without the left side response can be calculated.

Specifically, the term from the output entitled as 'count' is equivalent to cell 'a' of a 2x2 table (refer to illustration above). This enables the following calculations:

Total with response A (left side) = m1 =

count / (confidence / 100) = a / (confidence / 100)

Note: this is equal to the sum of cells a and b (a + b) from the 2x2 table.

Total with response A (left side) but not response B (right side) =

m1 – a

Note: this is equal to cell b from a 2x2 table.

Total without response A (left side) = m0 =

Total sample size (N) – m1

Note: this is equal to the sum of cells c and d (c + d) from the 2x2 table.

Total with response B (right side) = n1 =

Total sample size (N) * (expected confidence / 100)

Note: this is equal to the sum of cells a and c (a + c) from the 2x2 table.

Total without response B (right side) = n0 =

Total sample size (N) – n1

Note: this is equal to the sum of cells b and d (b + d) from the 2x2 table.

Total with response B (right side) but not response A (left side) =

n1 – c

Note: this is equal to cell b from a 2x2 table.

Having this information the ratio of the prevalence of response B among those with and without response A can be calculated as:

(a/m1) / (c/m0)

and the corresponding variance estimate as:

((m1-a)/(m1*a)) + ((m0-c)/(m0*c)).

The output that association rule discovery analysis provides uses lift as a measure of association. While the direction of an association is obvious via this measure, it provides no other intuitive information. The response ratio we present (RR), is analytically equivalent to a cumulative incident ratio, that is regularly used in closed cohort epidemiological analyses. This is therefore a convenient measure to use in association rule discovery for multiple reasons. Variance estimates and corresponding test of significance are easily produced without compromising information regarding the direction of association. When lift is equal to 1.0, so is the RR. However, by presenting the association in this manner we also can determine the difference in prevalence, as a percent, among those with response B and response A, as compared to those with response B, but not response A. A good example of this can be found in the appendix entitled "2-way associations – where left side of rule is deployed for Operation Desert Shield/Desert Storm". The second association rule presented here has a RR of 1.40, meaning that soldiers who reported that they deployed for Desert Shield/Desert Storm were 40% more likely to report that they experience a fair amount of stress due to separation from their family than did those soldiers who reported that they did not deploy for the same military operation. Conversely, the statistic, lift, provides no such information.

Technical Services: Data quality improvement for the Total Army Injury and Health Outcomes Database (TAIHOD)

Description:

The TAIHOD data warehouse has grown substantially over the past several years, and now contains many millions of records on close to 5 million current and former active duty Army soldiers. The information comprising the TAIHOD is collected from a number of different sources, each of which has different data representations as well as data quality characteristics. Examples of data sources with quality issues include Army personnel records, hazardous duty exposure records, disability records, outpatient visits, Health Risk Appraisal (HRA) records, and others. Assessing and improving upon the quality of the data contained within the TAIHOD is an important and time-consuming challenge, and one that begs for automated techniques to allow efficient and scientifically sound utilization of large databases.

On January 31, 2001, David Loshin from Knowledge Integrity met with members of the ARIEM/TAIHOD project team (LTC Paul Amoroso, Michelle Yore, MSPH, and Gary Schneider, MSPH) to discuss data analysis and data quality issues associated with the TAIHOD data warehouse. Critical data quality issues include:

1) Missing data
2) Incomplete records
3) Invalid information
4) Inconsistencies between different sources of data

At this meeting we discussed methodologies for discovery, characterization, definition, and management of data quality rules associated with the collection, preparation, and enhancement of large data sets. Documentation of methods used by Knowledge Integrity were also discussed, including some article reprints and a copy of David Loshin's new book, "Enterprise Knowledge Management – The Data Quality Approach." A preliminary demonstration of a new software product, *Guardian*, which is currently nearing alpha release, was also provided.

Scope

We believe that a small contract with Knowledge Integrity, Inc., will bring forth significant opportunities to make use of the proprietary data quality approach developed by David Loshin of Knowledge Integrity, Inc. Specifically, Knowledge Integrity's approach characterizes the cost associated with poor data quality, effectively measuring data quality as a function of conformance to user-defined data quality rules, and calculates the value associated with improving data quality. This approach is innovative and unique in the data quality industry.

The project goal will be to segment a particular data set, or small set of data sets, analyze the data for potential data quality rules, define those rules within their Guardian rule management system, and automate the generation of information validation applications to test and score the selected data sets. The choice of data sets will be made during the initial stages of the contract.

This contract will consist of 6-10 days of on-site consultation and training, as well as an alpha software license for the duration of the project with deployment of the Knowledge Integrity application software. The project goal is to characterize the most significant data quality problems and to specify recommendations to resolve those problems on a long-term scale. The fixed cost for this consultation, short-term software license, and deliverable report is $9500.00, broken out as:

Consultation, Temporary license	$4500.00
Assessment Report	$5000.00

Schedule

We expect that the time frame for deployment, training, and consultation should range from 3-6 months, depending on the visit schedule. Initial project deployment and training visits would be made more frequently, with consultation visits made less frequently. A start date will be set as soon as an agreement can be reached.

Deliverables

There are 3 deliverables to this project:

1) 6-10 days of on-site consultation and training.
2) A workable implementation of the software technology and methodologies culminating in a well-defined data validation framework for the selected data sets.
3) A report describing the process and delineating the value of a data quality approach to data fusion and enhancement within the TAIHOD environment.

Knowledge Integrity Incorporated
www.knowledge-integrity.com
516-505-3960
Contact: David Loshin
loshin@knowledge-integrity.com

Gender Misclassification in the TAIHOD database.

The goal of this project was to take an intuitively simple problem (to what degree and to what extent is gender misclassified in the TAIHOD?) and use the mining tools to model and discover important patterns among and between the various independent sources of gender codes within the database. In addition to providing a complex and vigorous exercise in the use of the data mining software, the expectation was that the results might also help us gain an understanding of a data quality issue with practical significance to our overall project. A high degree of certainty of an individuals true gender is critical for many analyses where gender is a covariate and also is important when matching records containing any significant degree of uncertainty as to the correct identity of the individual (an example might be the completion of an HRA survey by a family member who uses the sponsor's SSN on the form, therein creating some uncertainty as to who the data belongs to-- the spouse or the servicemember). Documentation of the SAS pilot project related to this effort is provided in a separate attachment.

PURPOSE

1) Implement SAS Enterprise Miner Software to develop a decision tree that will enable the identification and reclassification of persons with miscoded gender, or assign a gender to subjects with missing gender data fields in TAIHOD datasets. 2) Assess the performance of the model by estimating the extent of correct reclassification vs. incorrect reclassification via re-sampling methodology.

METHODS

Identifying Gender

We initially created a "master names dataset", the intent of which was to separately estimate the proportion of first and middle names that are associated with each gender. All individuals represented in the 1980, 1989, and 1998 DMDC files were used to create this dataset. First and middle names and suffixes were separately abstracted from all individuals represented in these files. The cumulative frequency of each unique name (e.g. Mike, John, Lauren, etc.) was calculated, this measure was also stratified by the gender coding in the DMDC files (gender coding from the DMDC files were assumed correct). The proportion of each name by gender was then calculated and this proportion was adjusted to account for the overrepresentation of males in the U.S. Army. This was done by taking dividing the ratio of males and females in the Army by the ratio of the same for any given name. The resulting estimates were representative of the likelihood of being a male or a female based on how common a name is (or is not) among individuals of each gender. This was done separately for each unique first name and for each unique middle name.

Using a similar approach, we also explored occupational and demographic data to enhance our accuracy in correctly identifying gender. The first 3 digits of each person's Military Occupational Specialty (MOS) code as well as data on each soldier's height, represented in inches, were used. Thus separate estimates of the proportion of each gender with each unique MOS, as well as height measured to the nearest inch, were calculated. Unfortunately, complete body weight data was unavailable for analysis.

Creating "Gold Standard" for Comparison to Predicted Model Results

A gender "gold standard" database was created concurrently. Three criteria had to be met for an individual to be included in the "gender gold standard" database. First, a gender specific ICD-9-CM diagnostic code from hospital records and/or gender specific responses to questions in the Army Health Risk Appraisal (HRA) between 1995 and 1998 were identified; these gender specific fields had to correspond to the gender coding in the same database. Second, multiple personnel data from the Defense Manpower Data Center (DMDC) records during the same years had to be available on these individuals and the gender coding in these files had to correspond to the gender coding in the hospital/HRA files with the gender specific responses. Lastly, no discrepancies between all available DMDC files in which an individual appeared were permitted.

The information regarding the prevalence of names (first and middle), MOS's and height among the genders was merged with the gender gold standard data on an individual basis. The resulting database therefore consisted of an individual's unique identifier, their gender (as determined via the "gold standard"), and continuous measures (between 0 and 1, inclusive) representing the prevalence of their first name, middle name, MOS, and height among each gender as determined from the "master names dataset." The Decision Tree node of SAS Enterprise Miner software was used to estimate the prevalence of these predictors and to ultimately predict gender. Examination of a sample of the initial decision tree results revealed that individuals with three names (not including their family name) were commonly misclassified. We hypothesized that this "extra" name would fall into two primary categories; females that hyphenated their family name after marriage (hyphens are not included in the DMDC names data) and males with a title following their name such as Sr., Jr., III, etc. Two indicator variables were therefore created, one representing the presence/absence of an "extra" name, and the other indicating if this name was one of the tiles predominately belonging to males. These indicator variables were added to the list of potential predictor variables that would be used by the decision tree to determine gender.

After developing the decision tree model by using the gender gold standard, we applied it to TAIHOD data sources; Hospitalizations (PASBA), Safety (ASMIS), Health Risk Appraisals (HRA) and the Ambulatory Data System (ADS) in which we were uncertain of the amount of underlying gender misclassification. This was applied only to the 1995 through 1998 files. ADS data was only available for 1998. Contingency tables were created representative of all possible "from gender/to gender" combinations, where the "from gender" axis could take on the values of 'male', 'female' or 'unknown' and the "to gender" axis could take on only the outcomes of 'male' or 'female'.

The discordant pairs from these contingency tables would not be informative without some baseline information regarding the actual underlying gender misclassification from each data-source, as it is possible that the decision tree reclassification could vary from the actual gender misclassification. Additionally, the overall quantity of reclassification is not of primary importance; the ratio of correct versus incorrect reclassification is however paramount in assessing the utility of the decision tree model.

Calculating the Underlying Rate of Gender Misclassification

The underlying misclassification of DMDC files was initially assessed. The DMDC files are updated on six-month intervals; therefore soldiers will be included in multiple DMDC files during their time in service. In some cases, the code for gender changes over time (e.g., a "male" soldier may get inadvertently coded as female during one or more of the DMDC files, or vice-versa). During the interval from 1980 through 1998 all changes in gender coding from successive pairings of DMDC files for all active duty service members were identified. The proportion calculated as the annual sum of these discrepant pairings divided by the size of the corresponding annual cohort of active duty Army personnel was used as an estimate of the underlying gender misclassification. The quantity of gender disagreement between the PASBA, ASMIS, HRA and ADS databases as compared to the DMDC file of the closest time proximity were used as estimates of the underlying gender misclassification from each of these data sources.

Validating Predictive Model Using Gold Standard

We employed a re-sampling technique to determine the quality of the gender reclassification. The aforementioned gender gold standard database was used for this exercise. Due to the three criteria that had to be met for inclusion into this database, there was virtually no doubt regarding the gender of these individuals. However, we knowingly reassigned gender to the incorrect value on a random 1.0% of this database. This was done 1000 times. For each of these 1000 re-samples, gender was reassigned on a random 1.0% and the previously constructed decision tree was applied. Since we know

with virtual certainty what the actual gender of each individual was, we were able to compare the predicted gender via the decision tree to the actual gender, and assess the quantity of correct and incorrect reclassification. Correct reclassification occurred when the decision tree reassigned the miscoded gender back to the gender gold standard value. Conversely, when gender was reassigned to the gender that did not agree with the gender gold standard data, incorrect reclassification occurred. We hypothesized that different sub-groups would be more prone to incorrect reclassification and/or less prone to correct reclassification. To examine this, the re-sampling technique was stratified by gender, and by the decision tree's predicted probability of being coded as a specific gender.

PRELIMINARY RESULTS

The cell counts and corresponding proportions of concordant and discordant pairs between the 'from gender' and 'to gender' axis for each data source that the decision tree model was applied to is displayed in Table 1 for the years 1995 through 1998 ('98 only for ADS). The sum of the percentages representing the four possible discordant pairings are 0.97%, 0.76%, 0.63% and 0.72% for the PASBA, ASMIS, HRA and ADS databases, respectively. Figure 1 represents the estimates of underlying gender misclassification, by year, for each TAIHOD data source used. The average gender misclassification during the 1995-1998 ('98 only for ADS) interval is 0.51%, 0.50%, 1.16% and 0.35% for the PASBA, ASMIS, HRA and ADS databases, respectively. Both the model reclassification and the underlying misclassification are minimal during the '95-'98 interval, however with the exception of the HRA data, the proportion of gender reclassification from the model exceeds the proportion of underlying misclassification. This suggests that a proportion of the gender reclassification is incorrect.

The results of the re-sampling analysis are presented in Table 2. The total number in each of the thousand re-samples, the mean number reclassified, as well as the mean numbers reclassified correctly and incorrectly, with corresponding percentages of the total are displayed. Additionally, the ratio between the correct and incorrect reclassification, with standard deviations and 95% confidence intervals, are presented.

For each resample the gold standard gender was deliberately miscoded on a random 1.0% of the subjects. However, the mean percentage (of the 1000 samples) of the total subjects reclassified exceeds 1.0% for the entire gold standard sample, as well as for each stratified analysis, providing evidence that there is underlying incorrect reclassification. The mean percent of the total reclassified correctly is equal to, or very close to, 1.0% for all the re-sampling analysis other than the two stratum representing a predicted

probability of 50-90% of being a specific gender, where this value was 0.82% and 0.62% for males and females, respectively. This provides evidence that the decision tree is likely to identify and reclassify gender on the vast majority of individuals that are incorrectly coded, with diminishing performance as the models certainty of classification decreases. The mean percent of the total reclassified incorrectly however varies greatly, ranging from 0.1% for the 'Pr (male) > 90%' strata to nearly 37.5% for the 'Pr (female) = 50-90%' strata. Large values of this measure suggest that an abundance of gender is being incorrectly reclassified. When the mean quantity (or corresponding percent of total) of correct reclassification exceeds the mean quantity of incorrect reclassification from the 1000 samples, good model performance is achieved. Conversely, when the quantity of incorrect reclassification exceeds the quantity of correct reclassification, the model performed poorly.

The mean ratio from the 1000 samples of correct versus incorrect reclassification for the 'all subjects' analysis was 2.41, meaning there were approximately 2.4 correct gender reclassifications for every 1.0 incorrect reclassification in this hypothetical situation where 1.0% misclassification is anticipated. Thus the decision tree would have reduced the overall quantity of gender misclassification. Examination of the stratified analyses show that this favorable reclassification was achieved for all males (ratio = 5.07) but not all females (ratio = 0.68). The decision tree model yielded favorable reclassification for individuals in the 'Pr (male) > 90%' stratum (ratio = 10.38) and for the 'Pr (female) > 90%' stratum (ratio = 1.44) than for the 'Pr (male) = 50-90%' and the 'Pr (female) = 50-90%' strata, which had corresponding correct/incorrect reclassification ratios of 0.06 and 0.02, respectively. Also noteworthy is the fact that the stratum representing the probability of 50%-90% of being coded as a specific gender, consisted of only 0.7% of all males and 2.1% of all females, however were responsible for 51.5% and 53.6% of all incorrect reclassification, respectively.

DISCUSSION

The estimates of underlying gender misclassification from each TAIHOD database assessed were quite low (refer to Fig 1), only the ASMIS and HRA data sources exceeded 1.0% at any time during the nineteen year interval spanning from 1980 to 1998. The ASMIS misclassification peaked in the early 1990's and the HRA in the later 1990's. It is not surprising that these data sources appear to be more prone to gender misclassification, as these data are collected after the event of interest and by self-report, respectively. It is well documented in epidemiological literature that information bias is more prominent as time increases between an event of interest and the data collection regarding that same event. Self-reported data, where subjects fill in a survey form such as the HRA, are prone to miscoded information due to an increased potential for subjects

to mistakenly fill in incorrect data fields. The PASBA and ADS data sources are conversely originate from medical records that are completed by medical professionals, and as such are perhaps less prone to erroneous information. The DMDC data, although as expected had consistently low rates of gender disagreement between subsequent records, has undergone noticeable increases in the 1990's. These increases however seem to coincide with periods of high levels of deployment by the U.S. military, which perhaps increases the difficulty in keeping accurate information on soldiers, particularly those who are deployed.

More than 99.0% of the data consisted of concordant 'from gender / to gender' pairings for each data source. Concordant female pairings were more common in the PASBA and ASMIS data sources than in the ASMIS and HRA data sources. This is consistent with prior knowledge that female soldiers have higher rates of medical care utilization than do male soldiers and thus simply have more records available for analysis of gender code pairing.

As previously noted the quantity of gender reclassification by the decision tree is of secondary importance to an accurate assessment of the quality of performance: that is, correct reclassification versus incorrect reclassification. Our ability to identity a subset of individuals for whom virtual certainty regarding the accuracy of their gender was possible, not only was useful for developing the decision tree, but also was of utmost importance to assess the quality of the gender reclassification. The deliberate recoding of gender to the incorrect value on a random 1.0% of this data was done to simulate a realistic gender misclassification scenario. A perfectly performing model would recode this 1.0% back to the correct value and would not reclassify the gender of any other subjects. By conducting this simulation many times (n=1000), where the 1.0% of the miscoded genders was randomly chosen from each simulation, we were able to provide estimates regarding both the models efficiency and its consistency over repeated trials. Furthermore, by employing this technique to specific subsets of the gender gold standard dataset, we were able to identify variation in the models performance in regards to both gold standard gender coding and the models certainty regarding gender classification.

Add in discussion regarding:

1) Better performance with more common names (large n = increased precision of estimates)

2) Potential role of ethnicity and its connection to #1

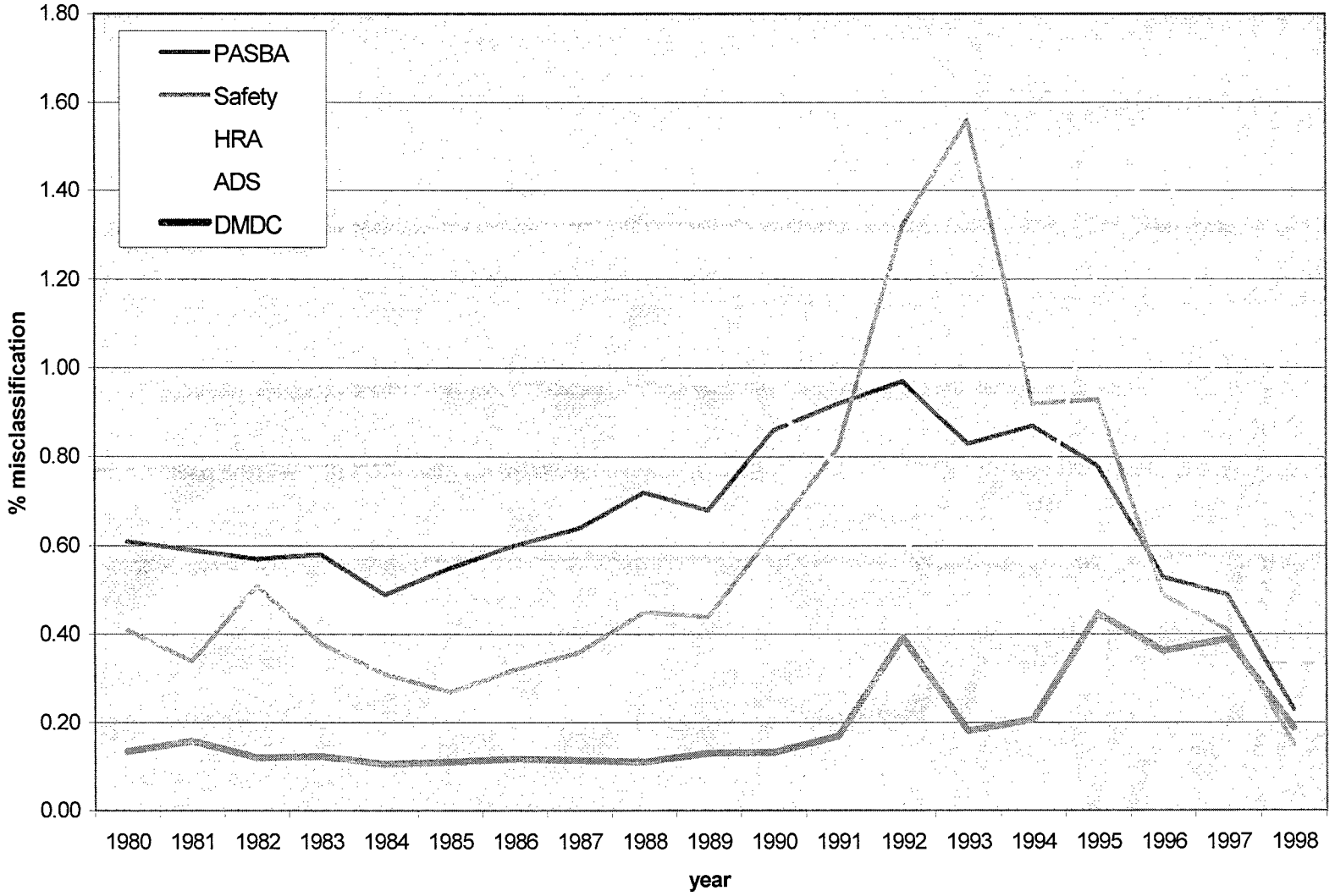
3) Better performance with Males vs. females (again, large n = increased precision of estimates)

4) Reformulation of re-sampling technique so that each simulation can be conducted with an equal sample size increasing the comparability of different strata. (yet again, large n = increased precision of estimates)

5) generalizability of methodology

[This report is incomplete.]

Figure 1. Gender misclassification of TAIHOD data sources as compared to DMDC data by year



Data Mining Of Data From TAIHOD Databases for the Purpose of Predicting Gender Where Gender is Misclassified

Prepared for: **US Army Research Institute of Environmental Medicine**
Attn: **Dr. Paul Amoroso, Michelle Yore**
Date: **January 12, 2001**

Project Summary

The goal of the data mining pilot as outlined in the statement of work was to identify data for the analysis, determine the data structure conducive to data mining and develop a working prototype that exhibits the core functionality of a data mining process flow using supervised modeling techniques. Inherent in the goal and stated in the statement of work was the commitment to transfer knowledge of the data mining process and the functionality of SAS Enterprise Miner.

The purpose of this summary document is to provide the US Army Research Institute of Environmental Medicine (USARIEM) some general pilot documentation for future reference.

The first step in the data mining process was to focus on a specific business problem that could be addressed and hopefully answered with data mining techniques. The problem that was identified was, how can data mining techniques help in predicting correct gender where gender was either missing or misclassified. The next step was to determine what data sources contained information that would help us build a predictive model. Since we focused on a predictive modeling process we needed to have an outcome or a way to correctly classify gender into a target variable. There are many different ways we could create a supervised model that would lend itself to predicting gender. The method we chose was to focus on data that contained gender related diagnoses or self-reported gender data in the HRA and hospital data sets. The idea behind this method was to identify a population in the data that we were relatively certain to be either male or female and indicate the population with a flag that would then become our target variable.

We discovered an inherent pitfall with this method. We are building a model that will predict gender from very narrow criteria. If a person without a gender assignment is scored with our model and they do not have any gender specific diagnoses or self-reported gender information than our model does not help in predicting their gender. If however, there are fields with data that are associated with gender that we have not used to build the flag, then our model might be helpful. Examining competing splits in a decision tree would be one way to determine what other variables might have been considered.

Some considerations that we explored in building the process flows were:

- Exploring data interactively (Insight node)
- How to partition data into training and validation data sets (Data Partition node)
- How to handle missing values (Replacement node)
- Getting rid of outliers (Filter Outliers node)
- Creating categorical nominal variables out of continuous interval variables (Transform Variables node)
- Determining what variables are associated with the target using R-squared (Variable Selection node)
- Model output and comparison (Assessment node)
- Applying new data to the model of choice (Score node)
- Understand scored data set (Code node)

In summary, the process of building this model was to choose an outcome that was specific to a successful assignment of gender and identify it as our target. The model will determine which fields from those that we used to compose our flag will contribute most to predicting gender and will then lend itself to giving predictive value to a new data set where gender is missing or possibly incorrectly assigned.

We also explored some unsupervised modeling techniques using cluster analysis. This method was helpful in finding patterns in the data that could be used to better understand how fields and their values were related to gender. An exercise that we talked about but didn't explore fully was taking the results of a cluster analysis and designating a cluster as the target that could be used in a supervised model. Another method for understanding the rules of a cluster is to select a cluster in the partitions tab of the cluster results and right click to open up "cluster profile." This will give a tree representation of the cluster and the rules associated with it.

The final model we built used Gary's proportion of gender by first name and middle name from the DMDC file. This model proved to be very accurate in predicting gender based solely on first and middle names.

We went though the exercise of taking the score code from a model such as Gary's and implementing it into a data warehouse through a user exit. We also explored using Code nodes to take our scored output and pull out records that fell above or below a certain probability cutoff.

A final note. The help documentation that comes with the software has much information on how to successfully utilize each node and some methodology on when and why to use nodes in a process flow.

Description of data preparation and progress on text mining portion of TATRC Information Mining Grant (USARIEM): Development of automated machine methods to analyze Parachute data in replication of previous published work of Amoroso, et al.

Submitted by David Jensen, PhD. UMass Computer Science Department and edited by LTC Amoroso.
2001.05.14

To assess the ability of automatic methods of free text analysis we begin with a previously defined problem (with a known solution) and attempt to replicate the prior results (in far less time) before attempting to apply the methods to larger and more complex datasets. Several years ago Amoroso et al published a study in which several thousand narrative reports of parachuting injuries were read in order to ascertain what the cause of injury was during a parachute jump. The results of these analyses were provided to us and will be used to train our text recognition software. Once trained, the model produced will be used on much larger datasets to test the effectiveness of the methods. The goal is to have an algorithm accurate enough to be useful and automatic enough to be run in a tiny fraction of the time required to do the same analysis "by hand".

This document briefly describes three steps used to prepare the parachute data for further analysis using data mining techniques. Such preparation is a large and important part of the knowledge discovery process (Brodley & Smyth 1997; Fayyad, Piatetsky-Shapiro, & Smyth 1996). Specifically, we applied a three-step process to prepare the parachute data for analysis: 1) variable selection, 2) text anonymization, and 3) text variable extraction. The latter two processes involved developing novel software or using software previously developed at the University of Massachusetts Center for Intelligent Information Retrieval.

Variable Selection

The data originally extracted for analysis contained 575 cases and 170 variables. The initial decision to include the variable in the extracted data set was based on the meaning of the variable, not on its intrinsic suitability for analysis.

We analyzed the distribution of each variable independently, to determine its suitability for analysis. In this process, we accounted for the data mining algorithms we intended to apply later (e.g., algorithms for constructing decision trees and association rules) and the data requirements of those algorithms.

Of the original 170 variables, 112 had values for fewer than 100 cases. Of those, the vast majority had values in fewer than 40 cases. For example, values for the variable AGE was present in only 3 cases and the

availability and use of goggles was known in only 3 cases. Such variables rarely provide substantial predictive power, and the algorithms we intend to use are not highly tolerant of missing values. All of these variables were eliminated.

Of the remaining variables, 18 had essentially only one value. Without substantial variation in the values of the variable, there is little point in including it in the analysis. These variables were eliminated.

After eliminating these two classes of variables, 40 variables remain, including one that is an ID variable. Four of the remaining variables appeared essentially useless for our purposes of analysis. The variables denoted the sequence number of accident, the files source, the date, and whether the case was a member of duplicate set. These variables were eliminated. In addition, the variables TIME and HOUR were essentially identical, so HOUR was eliminated.

Six of the remaining variables weren't really useful for prediction. Instead, they were useful for constructing a "dependent" variable (what a statistical model is attempting to predict). The six variables are:

Name	Annotation
DAYHOSP	Days hospitalized
DAYLOST	Days lost
DAYREST	Days restricted
INJCOST	Total injury cost
NBPART1	Body part 1 affected
PINJCOST	Injury cost this person
Primary_Error01	Parachutist error 1
Primary_Error02	Parachutist error 2

These variables were eliminated from the set of predictors (but will be used later to produce alternative dependent variables).

Two of the remaining variables encoded complex data which required recoding. Specifically, GRADE and MOS1T4. ARIEM staff provided procedures to recode the variables into discrete categories that are more amenable to analysis.

Text Anonymization

In addition to numeric and symbolic variables (e.g., age and rank, respectively), the parachute data contains text fields with narrative

descriptions of the accident. One of the goals of the project was to use data derived from the text fields as part of the analysis. However, much of the textual data contains personal names and social security numbers (SSNs) of military personnel. Such personal identifiers are not necessary for the analysis, but they make it impossible for outside researchers (e.g., UMass) to work on the data without producing legitimate privacy concerns.

To alleviate such concerns, we wrote a program to automatically anonymize personal identifiers (names and SSNs) by converting them to surrogate names and numbers. The program accepts a text string, identifies each proper name and SSN, creates a substitute name or SSN for each, carries out the substitution, and then stores the anonymized text string. This substitution process preserves repeated uses of the same name (called "co-reference" in the field of natural language processing), so the text remains meaningful to human readers.

This process was made more difficult because the text strings in the parachute data were stored in all capital letters, rather than upper- and lower-case letters which would have provided capitalization cues for proper names. However, the process was made easier by the rank identifications that often precede the first use of a proper name (e.g., PFC Smith). As a side benefit, this process converted the text to upper- and lower-case, to aid human readers.

The program went through several iterations where changes were made to the software at UMass, UMass personnel tested the program on ten text strings that had been hand-anonymized by USARIEM personnel, and then the program was tested on the full data at USARIEM. This process allowed the software to be developed at UMass without compromising the privacy of military medical records.

Text Variable Extraction

A large number of medical databases, including the parachute data, contain text fields that provide narrative descriptions of accidents or medical conditions. However, essentially all data mining techniques assume that cases are represented by numeric and symbolic variables. Thus, one of the first steps in using text fields is to recode key aspects of the meaning of a narrative description into one or more numeric or symbolic variables.

The meaning of narrative descriptions are generally easy for humans to understand, but human language has proven surprisingly resistant to automated processing, despite several decades of work by computer

scientists and linguists. That said, several relatively simple approaches based on statistical techniques have proven successful for limited tasks. In addition, "tagging" techniques that identify the linguistic structure of sentences has also been developed and successfully applied.

Our approach to extracting variables from text uses both parsing and some limited statistical approaches. Specifically, we use JTAG, a tagger that predicts the part-of-speech for each word in the sentence. JTAG was developed at UMass and has been used in a variety of information retrieval and organization systems.

After words have been associated with a particular part of speech, then useful variables are created based on frequently occurring nouns, verbs, and noun phrases. Specifically, if a word or phrase occurs in more than 10% of the records and less than 90%, it becomes the basis for creating a variable. Each variable created from the text fields is Boolean. That is, they have a value of 1 if the word or phrase occurs in that case's text field and they have a value of 0 if the word or phrase does not occur in the text field.

Results of the initial analyses are pending.

References

Brodley, C. E. and Smyth, P. (1997). Applying classification algorithms in practice. Statistics and Computing 7:45-56.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). From data mining to knowledge discovery in databases. AI Magazine. Fall. 37-54.

USARIEM
Data Warehouse Pilot Summary Document
January 15, 2001

Introduction

SAS Institute Inc ('SAS') has had correspondence with LTC. Paul Amoroso and Michelle Yore of the United States Army Research Institute of Environmental Medicine ('USARIEM') to discuss and evaluate the scope and feasibility of a data warehouse pilot project.

From this discussion, USARIEM outlined the structure, and goals of a data warehouse system for building reporting datasets and a data mart for use by Enterprise Miner. The goal and end result of this pilot will be a working system that will lay the foundation for a Data Management system. This document presents a summary of the work SAS provided surrounding the proposed pilot project.

Period Of Performance

The work for this pilot was performed over a total of five days. December 6-7 2000 and January 10-12, 2001. SAS performed the work on-site at USARIEM in Natick, MA in cooperation with selected USARIEM staff.

Project Summary

USARIEM currently receives 8 datasets from various sources (DMDC, etc.) in either text or MS Access formats. The files are received at different intervals (bi annually, annually, etc.) and data processing is done to incorporate the new information into the historical datasets.
What USARIEM was to be able to do was to use WA to administer the data updates and new data file creation for end user access and data mining.

Critical success factors included: a working prototype that provides core functionality for data administration and new file creation, and an understanding of WA tools by USARIEM personnel so further enhancements can continue.

Work Performed

The following tasks were accomplished during the period of performance.

Task 1

SAS personnel configured the Warehouse environment to make best use of the USARIEM systems. Warehouse administrator was configured on a Windows NT workstation, to provide development and enhancement capabilities. The resulting data repositories were written to a Windows NT data server ('HP RACK'). The processes for repository production were configured to process on the HP RACK to take advantage of its computing power and the network configuration. The appropriate SAS/CONNECT scripts were used.

Task 2
The pilot team achieved transparent access to Safety data stored in MS Access databases intermittently supplied to USARIEM.

Task3
The pilot team created process flows for 4 data areas of the TAIHOD databases.

- A 'skeleton' process was created to produce DMDC Master data sets, in which new data was created for every six-month period. USARIEM personnel decided to keep all data sets separate, because of the shear size.
- A process was created to produce a Safety dataset. The data was accessed from an MS Access database supplied to USARIEM. The process read the new data, transformed the data using existing USARIEM code, and appended the data to the existing data source.
- A process was created to produce a PASBA dataset. The data was accessed from 3 text files supplied containing Army, Air Force and Navy data. The process will append the new data to the existing data set, after duplicating the data set.
- A process was created to produce an HRA dataset. This data was accessed from a text file. The process will append the new data to the existing data set, after duplicating the data set.

The pilot team created sample process flows for 2 subject areas, in preparation for data mining.

- Safety-DMDC and PASBA-DMDC data sets and process flows were created. The ultimate use of these datasets is for data mining. These samples were created to familiarize USARIEM personnel with the objects and processes available for developing suitable data mining data sets.

Task 4
SAS provided knowledge transfer to appropriate USARIEM personnel after prototype development to ensure a smooth transition from development to implementation and ease of operations and maintenance.

Technical Configuration

Architecture
Server: HP Netserver running Windows NT 4.0
Workstation: Miscellaneous, running Windows NT.

Data
SAS v8 datasets on HP Windows NT 4.0 data server.

Display
PC-based GUI provided by out of the box SAS/WA application.

Installation
The finished WA prototype developed and tested on selected PC.
Required SAS Product Configuration on the NT Server (SAS v8)
Base SAS
SAS CONNECT

SAS ACCESS to ODBC

Required SAS Product Configuration on miscellaneous PCs
Base SAS
SAS CONNECT
SAS Warehouse Administrator